JP9200601324S1

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-259515
(43)Date of publication of application : 24.09.1999

(51)Int.CI. G06F 17/30
G06F 17/27
G06F 17/21

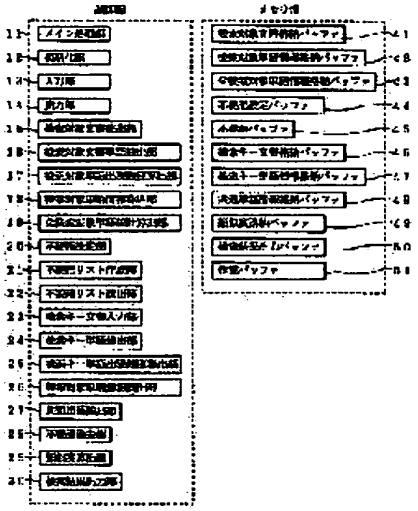
(21)Application number : 10-061726
(22)Date of filing : 12.03.1998

(71)Applicant : TOSHIBA CORP
TOSHIBA COMPUT ENG CORP

(72)Inventor : TANOSAKI YASUO
NAKAMOTO YUKIO
NISHINA TAKUYA
KUBOTA NAOHIDE

(54) SIMILAR DOCUMENT RETRIEVAL DEVICE AND METHOD AND RECORDING MEDIUM RECORDING SIMILAR DOCUMENT RETRIEVAL PROGRAM

(57)Abstract:
PROBLEM TO BE SOLVED: To improve both inter-document similarity calculation accuracy and similar document retrieval accuracy via the optimization of a list of unnecessary words by deciding some of extracted words as unnecessary words, deleting the unnecessary words from a retrieval key document and a retrieval object document and calculating the similarity between both documents.
SOLUTION: Some of words extracted by a word extraction means are decided as unnecessary words based on the occurrence frequency of each designated unnecessary word. Then the unnecessary words are deleted from a retrieval key document and a retrieval object document, and the similarity is calculated between both documents. An unnecessary word deletion part 28 of this similar document retrieval device deletes the words equivalent to the unnecessary words stored in an unnecessary word buffer 45 from a retrieval keyword information storing buffer 47 and a retrieval object word information storing buffer 42. A similarity calculation part 29 calculates the similarity between the retrieval key document and the retrieval object document based on the information which are stored in the buffer 47, the buffer 42 and a common word information storing buffer 48.



LEGAL STATUS
[Date of request for examination]
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－259515

(43)公開日　平成11年(1999)９月24日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ	17/30	Ｇ０６Ｆ	15/40	３７０Ａ
	17/27		15/20	５５０Ａ
	17/21			５７０Ｎ
			15/403	３５０Ｃ

審査請求　未請求　請求項の数７　ＯＬ　（全 12 頁）

(21)出願番号　特願平10－61726

(22)出願日　平成10年(1998)３月12日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(71)出願人　000221052
東芝コンピュータエンジニアリング株式会社
東京都青梅市新町３丁目３番地の１

(72)発明者　田野崎　康雄
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内

(74)代理人　弁理士　須山　佐一

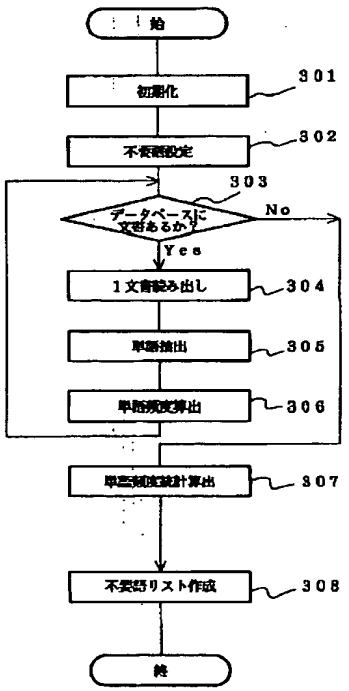
最終頁に続く

(54)【発明の名称】　類似文書検索装置、類似文書検索方法、および類似文書検索のためのプログラムが記録された記録媒体

(57)【要約】

【課題】　従来、文書間の類似度の算出において、各文書内の単語群から、類似度算出において不要と思われる単語（不要語）をユーザ自身が一つ一つ選択している。このためユーザの負担が大きく不要語の選択漏れも発生する確率が高い。

【解決手段】　文書間の類似度算出において各文書データから抽出される単語群の中から排除すべき種類の単語（不要語）を設定するための基準の設定を行う不要語設定部２０と、このバッファ４４内の設定内容に基づいて不要語リストを自動作成する不要語リスト作成部２１とを備える。基準の設定は、例えば、任意の単語（不要語）をユーザが１つ乃至複数指定することによって行われ、この場合、不要語リスト作成部２１は、指定された不要語の最小出現頻度を基準値として求め、文書データより抽出された単語群のうち、算出された出現頻度が基準値以上のすべての単語を不要語とする。



1

【特許請求の範囲】

【請求項１】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、
前記検索キー文書および前記検索対象文書を含む複数の文書データが格納された文書データ格納手段と、
任意の単語を指定する単語指定手段と、
前記文書データ格納手段に格納された各文書データから単語を抽出する単語抽出手段と、
前記単語指定手段により指定された任意の単語および前記単語抽出手段により抽出された単語の前記各文書データ中での出現頻度をそれぞれ算出する出現頻度算出手段と、
前記出現頻度算出手段によって算出された前記任意の単語の出現頻度を基準として、前記単語抽出手段により抽出された単語のうちの少なくとも一部の単語を不要語として判定する不要語判定手段と、
前記検索キー文書および前記検索対象文書から前記不要語判定手段により判定された不要語をそれぞれ除いて両文書間の類似度を算出する手段と、
を具備することを特徴とする類似文書検索装置。

【請求項２】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、
前記検索キー文書および前記検索対象文書を含む複数の文書データが格納された文書データ格納手段と、
任意の単語を指定する単語指定手段と、
前記文書データ格納手段に格納された各文書データから単語を抽出する単語抽出手段と、
前記単語指定手段により指定された任意の単語および前記単語抽出手段により抽出された単語の前記各文書データ中での出現頻度をそれぞれ算出する出現頻度算出手段と、
前記単語抽出手段により抽出された単語のうち、前記出現頻度算出手段によって算出された出現頻度が、前記任意の単語について算出された出現頻度以上の単語を不要語として判定する不要語判定手段と、
前記検索キー文書および前記検索対象文書から前記不要語判定手段により判定された不要語をそれぞれ除いて両文書間の類似度を算出する手段と、
を具備することを特徴とする類似文書検索装置。

【請求項３】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、
前記検索キー文書および前記検索対象文書を含む複数の文書データが格納された文書データ格納手段と、
複数の任意の単語を指定する単語指定手段と、
前記文書データ格納手段に格納された各文書データから単語を抽出する単語抽出手段と、
前記単語指定手段により指定された複数の任意の単語および前記単語抽出手段により抽出された単語の前記各文書データ中での出現頻度をそれぞれ算出する出現頻度算出手段と、
前記単語抽出手段により抽出された単語のうち、前記出現頻度算出手段によって算出された出現頻度が、前記出現頻度算出手段によって算出された複数の任意の単語の出現頻度のうちの最小出現頻度以上の単語を不要語として判定する不要語判定手段と、
前記検索キー文書および前記検索対象文書から前記不要語判定手段により判定された不要語をそれぞれ除いて両文書間の類似度を算出する手段と、
を具備することを特徴とする類似文書検索装置。

【請求項４】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、
前記検索キー文書および前記検索対象文書を含む複数の文書データが格納された文書データ格納手段と、
前記文書データ格納手段に格納された各文書データから単語を抽出する単語抽出手段と、
前記単語抽出手段により抽出される単語のうち不要語とすべき単語の数を任意に指定する不要語数指定手段と、
前記単語抽出手段により抽出された単語の前記各文書データ中での出現頻度を算出する出現頻度算出手段と、
前記単語抽出手段により抽出された単語のうち、前記出現頻度算出手段により算出された出現頻度が高いものから優先に、前記不要語数指定手段により指定された数の単語を不要語として判定する不要語判定手段と、
前記検索キー文書および前記検索対象文書から前記不要語判定手段により判定された不要語を除いて両文書間の類似度を算出する手段と、
を具備することを特徴とする類似文書検索装置。

【請求項５】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、
前記検索キー文書および前記検索対象文書を含む複数の文書データが格納された文書データ格納手段と、
前記文書データ格納手段に格納された各文書データから単語を抽出する単語抽出手段と、
前記単語抽出手段により抽出された単語の前記各文書データ中での出現頻度を算出する出現頻度算出手段と、
任意の出現頻度を指定する出現頻度指定手段と、
前記単語抽出手段により抽出された単語のうち、前記出現頻度算出手段によって算出された出現頻度が、前記出現頻度指定手段により指定された任意の出現頻度以上の単語を不要語として判定する不要語判定手段と、
前記検索キー文書および前記検索対象文書から前記不要語判定手段により判定された不要語を除いて各文書間の類似度を算出する手段と、
を具備することを特徴とする類似文書検索装置。

【請求項６】　ある文書を検索キー文書としてこの検索

キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索方法において、
任意の単語を指定する工程と、
前記検索キー文書および前記検索対象文書を含む複数の文書データから単語を抽出する工程と、
前記指定された任意の単語および前記抽出された単語の前記各文書データ中での出現頻度をそれぞれ算出する工程と、
前記算出された前記任意の単語の出現頻度を基準として、前記抽出された単語のうちの少なくとも一部の単語を不要語として判定する工程と、
前記検索キー文書および前記検索対象文書から前記判定された不要語をそれぞれ除いて両文書間の類似度を算出する工程とを有することを特徴とする類似文書検索方法。

【請求項７】　ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索するためのプログラムが記録された記録媒体において、
任意の単語を指定する単語指定手段と、
前記検索キー文書および前記検索対象文書を含む複数の文書データから単語を抽出する単語抽出手段と、
前記単語指定手段により指定された任意の単語および前記単語抽出手段により抽出された単語の前記各文書データ中での出現頻度をそれぞれ算出する出現頻度算出手段と、
前記出現頻度算出手段によって算出された前記任意の単語の出現頻度を基準として、前記単語抽出手段により抽出された単語のうちの少なくとも一部の単語を不要語として判定する不要語判定手段と、
前記検索キー文書および前記検索対象文書から前記不要語判定手段により判定された不要語をそれぞれ除いて両文書間の類似度を算出する手段とを有するプログラムが記録されていることを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書データベースから、文書間の類似度に基づく文書データの検索を行う類似文書検索装置、類似文書検索方法、および類似文書検索のためのプログラムが記録された記録媒体に関する。

【０００２】

【従来の技術】近年、大量の電子化された文書データが流通するようになり、自動分類等を行う目的で、文書データベース中から指定された文書（以下、検索キー文書と呼ぶ）に類似する文書の自動検索を行うシステムが実用化されてきている。この文書検索システムでは、検索キー文書に含まれている単語と検索対象となる文書（以下、検索対象文書と呼ぶ）に含まれている単語とを比較し、共通する単語の種類、出現場所、出現回数などからベクトル空間法により類似度を算出し、類似度の高い検索対象文書を検索結果として出力する。

【０００３】このとき、類似文書検索を行う上で不要な単語（文書の内容を特徴付けるものではない一般的な単語）を含めた類似度の算出は検索精度を落とす原因となり得ることから、予め不要語リストを作成しておき、文書から単語を抽出する際に不要語リストを参照して、不要語に相当する単語については文書から抽出しないようにする方法をとっている。

【０００４】しかしながら、通常、不要語リストの作成においては、不要語とすべき単語の種類をユーザが１つ１つ決定する必要があり、しかも検索対象文書データベースの種類毎に別々の不要語リストを用意する必要がある。このような不要語リストの作成作業は、ユーザにとって大きな負担となるばかりか、不要語の選択の個人差によって、類似文書検索の精度に大きなばらつきが生じるという問題がある。

【０００５】

【発明が解決しようとする課題】このように、精度の高い類似文書検索を行うためには、文書から抽出すべき単語対象から不要語を排除することが好ましいが、そのためには検索対象文書データベースの種類毎に不要語リストを人手により作成する必要があり、ユーザに負担を強いることになる。また、不要語の選択漏れはもちろん、ユーザによる不要語の選択の個人差が検索結果に色濃く反映されてしまい、類似文書検索の精度のばらつきが生じやすいという問題がある。

【０００６】本発明はこのような課題を解決するためになされたもので、最適な不要語リストを自動的に作成でき、不要語リストの最適化による文書間の類似度算出精度の向上並びに類似文書検索精度の向上を図ることのできる類似文書検索装置、類似文書検索方法、および類似文書検索のためのプログラムが記録された記録媒体の提供を目的としている。

【０００７】

【課題を解決するための手段】上記した目的を達成するために、本発明の類似文書検索装置は、請求項１に記載されるように、ある文書を検索キー文書としてこの検索キー文書と類似する文書を複数の検索対象文書の中から検索する類似文書検索装置において、前記検索キー文書および前記検索対象文書を含む複数の文書データが格納された文書データ格納手段と、任意の単語を指定する単語指定手段と、前記文書データ格納手段に格納された各文書データから単語を抽出する単語抽出手段と、前記単語指定手段により指定された任意の単語および前記単語抽出手段により抽出された単語の前記各文書データ中での出現頻度をそれぞれ算出する出現頻度算出手段と、前記出現頻度算出手段によって算出された前記任意の単語の出現頻度を基準として、前記単語抽出手段により抽出された単語のうちの少なくとも一部の単語を不要語とし

て判定する不要語判定手段と、前記検索キー文書および前記検索対象文書から前記不要語判定手段により判定された不要語をそれぞれ除いて両文書間の類似度を算出する手段とを具備することを特徴とする。

【０００８】本発明においては、複数の文書データから抽出された単語群の中から、ユーザにより不要語の代表として任意に指定された単語に対して算出された出現頻度を基準として、単語抽出手段により抽出された単語のうちの少なくとも一部の単語を不要語として判定し、検索キー文書および検索対象文書から不要語をそれぞれ除いて両文書間の類似度を算出することによって類似文書検索を行う。

【０００９】例えば、請求項２に記載されるように、単語抽出手段により抽出された単語のうち、算出された出現頻度が、前記任意の単語について算出された出現頻度以上の単語を不要語として判定したり、或いは、請求項３に記載されるように、単語抽出手段により抽出された単語のうち、算出された出現頻度が、複数の任意の単語の出現頻度のうちの最小出現頻度以上の単語を不要語として判定する。更には、請求項４に記載されるように、単語抽出手段により抽出された単語のうち、算出された出現頻度が高いものから優先に予め指定された数の単語を不要語として判定したり、請求項５に記載されるように、単語抽出手段により抽出された単語のうち、算出された出現頻度が、予め指定された任意の出現頻度以上の単語を不要語として判定する。

【００１０】以上の発明により、文書データに含まれる単語群の中からの不要語の抽出を自動化できる。すなわち、ユーザは、例えば、代表的な不要語に当たる任意の単語を１つ乃至数個入力したり、任意の不要語の数を入力したり、基準の出現頻度を入力するだけで、希望するものに近い不要語リストを得ることができ、類似文書検索の全体的な効率を高めることができ、また、検索対象文書データベースの種類毎に最適かつ妥当な不要語を漏れなく迅速抽出することができるので、類似文書検索の精度の向上と安定化を図ることができる。

【００１１】また、複数の任意の単語の出現頻度のうちの最小出現頻度以上の単語を不要語として判定することにより、ユーザ毎の個人差が不要語のリストの違いに現れる度合が小さくなり、この点からも、類似文書検索の精度の向上と安定化を図ることができる。

【００１２】

【発明の実施の形態】以下、本発明の一実施例を図面を参照しながら説明する。

【００１３】図１は本発明の実施形態である類似文書検索装置のハードウェア構成を示す図である。同図に示すように、本実施形態の類似文書検索装置は、ＣＰＵ、メモリなどから構成される制御装置１、キーボードなどの入力装置２、類似文書検索の過程や結果などを表示する表示装置３、文書データや類似文書検索のために必要な各種データを格納する外部記憶装置４などから構成される。

【００１４】図２は本実施形態の類似文書検索装置の制御装置１の構成を示す機能ブロック図である。同図に示すように、制御装置１は制御部とメモリ部で構成される。

【００１５】制御部は、メイン処理部１１、初期化部１２、入力部１３、出力部１４、検索対象文書読出部１５、検索対象文書単語抽出部１６、検索対象単語出現頻度算出部１７、検索対象単語情報算出部１８、全検索対象単語統計算出部１９、不要語設定部２０、不要語リスト作成部２１、不要語リスト読出部２２、検索キー文書入力部２３、検索キー単語抽出部２４、検索キー単語出現頻度算出部２５、検索対象単語情報読出部２６、共通単語抽出部２７、不要語除去部２８、類似度算出部２９、検索結果出力部３０などで構成されている。

【００１６】メモリ部は、検索対象文書格納バッファ４１、検索対象単語情報格納バッファ４２、全検索対象単語情報格納バッファ４３、不要語設定バッファ４４、不要語バッファ４５、検索キー文書格納バッファ４６、検索キー単語情報格納バッファ４７、共通単語情報格納バッファ４８、類似度格納バッファ４９、検索結果出力バッファ５０、作業バッファ５１などで構成されている。

【００１７】初期化部１２は、各バッファ４１，４２，…，５１の初期化を行う。入力部１３は、入力装置２を通してユーザより入力されたデータを制御部に入力する。出力部１４は、制御部の出力データを表示装置３に出力する。

【００１８】検索対象文書読出部１５は、ユーザにより指定された検索対象文書を外部記憶装置４から読み込み、読み込んだ検索対象文書を検索対象文書格納バッファ４１に格納する。

【００１９】検索対象文書単語抽出部１６は、検索対象文書格納バッファ４１に格納された文書データから単語を切り出し、切り出された単語群の中からその文書の内容を特徴付ける単語を抽出し、抽出された単語種を検索対象単語情報格納バッファ４２に格納する。ここで、単語の切り出しは形態素解析等によって行われ、その文書の内容を特徴付ける単語の抽出は、単語の品詞情報に基づいて、例えば「名詞」や「サ変名詞」の単語を選択することによって行われる。

【００２０】検索対象単語出現頻度算出部１７は、検索対象単語情報格納バッファ４２に格納された個々の単語について抽出元文書内での出現頻度（出現数）を算出し、算出された出現頻度を検索対象単語情報格納バッファ４２に単語と対応付けて格納する。

【００２１】検索対象単語情報書込部１８は、検索対象単語情報格納バッファ４２に格納された各検索対象文書の単語情報と出現頻度の情報を読み出して外部記憶装置４に書き込む。

7

【００２２】全検索対象単語統計算出部１９は、外部記憶装置４に格納されている各検索対象文書の単語出現頻度の情報を、順次、読み出して検索対象文書格納バッファ４１に書き込み、全検索対象文書から抽出された単語の種類毎に、出現頻度の統計値例えば出現文書数などを算出し、その結果を全検索対象単語統計として全検索対象単語情報格納バッファ４３に格納する。

【００２３】不要語設定部２０は、文書間の類似度算出において各文書データから抽出される単語群の中から排除すべき種類の単語（不要語）を設定するための基準の設定をユーザより受け付けて、その設定された基準を不要語設定バッファ４４に格納する。このときの基準の設定方法には、任意の単語（不要語）を１つ乃至複数指定する方法、不要語の数を指定する方法、出現頻度の基準値を指定する方法がある。

【００２４】不要語リスト作成部２１は、全検索対象単語情報格納バッファ４３に格納されている全検索対象単語情報と不要語設定バッファ４４内の設定内容に基づいて不要語リストを作成し、作成された不要語リストを外部記憶装置４に格納する。

【００２５】この不要語リストの作成において、不要語設定部２０にて任意の不要語が１つ指定された場合は、その不要語について算出された出現頻度以上の単語を不要語として不要語バッファ４５に格納し、複数の不要語が指定された場合は、その不要語について算出された出現頻度のうちの最小出現頻度以上の単語を不要語として不要語バッファ４５に格納する。また、不要語設定部２０にて不要語の数が指定された場合は、全単語のうち出現頻度が高いものから優先に、指定された不要語数の単語を不要語として不要語バッファ４５に格納する。また、不要語設定部２０にて出現頻度の基準値が指定された場合は、全単語のうち指定された出現頻度の基準値以上の単語を不要語として不要語バッファ４５に格納する。

【００２６】不要語リスト読出部２２は、外部記憶装置４に格納されている不要語リストを読み込み、不要語バッファ４５に格納する。

【００２７】検索キー文書入力部２３は、入力装置２から入力された検索キー文書を検索キー文書格納バッファ４６に格納する。

【００２８】検索キー単語抽出部２４は、検索キー文書格納バッファ４６に格納された検索キー文書からの単語の切り出しを行い、切り出された単語群のなかから、その検索キー文書の内容を特徴付ける単語種を抽出し、抽出された単語種を検索キー単語情報格納バッファ４７に格納する。ここで、単語の切り出しは形態素解析等により行われ、文書の内容を特徴付ける単語の抽出は、単語の品詞情報に基づいて、例えば「名詞」や「サ変名詞」の単語を選択することによって行われる。

【００２９】検索キー単語出現頻度算出部２５は、検索

8

キー単語抽出部２４によって抽出された個々の単語について、抽出元文書内での出現頻度（出現数）を算出し、算出された出現頻度を検索キー単語情報格納バッファ４７に格納する。

【００３０】検索対象単語情報読出部２６は、外部記憶装置４に格納されている文書データベース中の各検索対象文書の単語情報とその出現頻度の情報を１文書毎に呼び出し、検索対象単語情報格納バッファ４２に格納する。

【００３１】共通単語抽出部２７は、検索キー単語情報格納バッファ４７および検索対象単語情報格納バッファ４２から検索キー文書および検索対象文書中に共通に存在する単語情報とその出現頻度の情報を読み出し、共通単語情報格納バッファ４８に格納する。

【００３２】不要語除去部２８は、検索キー単語情報格納バッファ４７および検索対象単語情報格納バッファ４２から、不要語バッファ４５に格納されている不要語に当たる単語を削除する。

【００３３】類似度算出部２９は、検索キー単語情報格納バッファ４７、検索対象単語情報格納バッファ４２および共通単語情報格納バッファ４８にそれぞれ格納された情報に基づき、ベクトル空間法等によって検索キー文書と検索対象文書との類似度を算出し、算出された類似度を類似度格納バッファ４９に格納する。

【００３４】検索結果出力部３０は、類似度格納バッファ４９に格納されている検索対象文書毎の類似度から、類似検索結果とする文書情報（例えば、文書ＩＤ）を検索結果出力バッファ５０に格納し、検索結果出力バッファ５０の内容を出力部１４を通じて表示装置３に出力する。

【００３５】次に、本実施形態の類似文書検索装置の動作を説明する。

【００３６】最初に、文書データベースおよび不要語リストを作成する動作について図３乃至図１０を参照して説明する。

【００３７】まず、初期化部１２が起動され、全バッファの初期化が行われる（ステップ３０１）。続いて、不要語設定部２０が起動され、不要語を設定するための基準の設定が行われる（ステップ３０２）。不要語を設定するための基準は、以下の３通りの方法の中からユーザにより任意に選択された方法で設定される。

【００３８】第１の方法は、ユーザが任意の数の単語（不要語）を指定し、この指定単語について算出された出現頻度（指定単語が複数の場合は各不要語について算出された出現頻度のうちの最小出現頻度）を基準値とし、そして文書データより抽出された単語群のうち、算出された出現頻度が基準値以上のすべての単語を不要語とする方法である。例えば、図４に示すように、文書を特徴付ける性質を持たない一般的な単語例えば「こと」「装置」などが指定され、これらの単語について算出さ

れた出現頻度のうち最小出現頻度を基準値として、出現頻度がこの基準値以上の単語を不要語とする。

【００３９】第２の方法は、ユーザが不要語の数（或いは出現頻度値の順位）を任意に指定し、文書データから抽出された単語群のうち、算出された出現頻度が高いものから優先に、前記指定された数の単語を不要語とする方法である。例えば、図５に示すように、「指定順位＝２（不要語数＝２）」のように指定された場合、算出された出現頻度が上位２位までの単語を不要語とする。

【００４０】第３の方法は、ユーザが出現頻度の基準値を任意に指定し、文書データから抽出された単語群のうち、算出された出現頻度が前記指定された出現頻度の基準値以上のすべての単語を不要語とする方法である。例えば、図６に示すように、「指定出現頻度＝５００以上」にように指定された場合、出現頻度が５００以上のすべての単語を不要語とする。

【００４１】これら３つの方法のいずれかによって設定された不要語の設定基準は不要語設定バッファ４４に格納される。

【００４２】次に、検索対象文書読出部１５が起動される。検索対象文書読出部１５は外部記憶装置４にまだ処理を終えてない検索対象文書があるか否かを判断し（ステップ３０３）、検索対象文書があれば、図７に示すように、その検索対象文書を検索対象文書格納バッファ４１に格納する（ステップ３０４）。

【００４３】次に、検索対象文書単語抽出部１６が起動される。検索対象文書単語抽出部１６は、検索対象文書格納バッファ４１に格納された検索対象文書から形態素解析等によって単語を切り出し、切り出された単語群から「名詞」や「サ変名詞」などの文書の内容を特徴付ける単語を抽出し、抽出された単語を検索対象単語情報格納バッファ４２に格納する（ステップ３０５）。

【００４４】続いて、検索対象単語出現頻度算出部１７が起動される。検索対象単語出現頻度算出部１７は、検索対象単語情報格納バッファ４２に格納されている個々の単語について、その抽出元文書中での出現頻度をそれぞれ算出し、例えば図８に示すように、算出された出現頻度の情報を単語と対応付けて検索対象単語情報格納バッファ４２に格納する。以降、この単語と出現頻度の情報を「単語情報」と呼ぶ。なお、図８において、「文書」という単語に対応して記述された「頻度２」は「文書」という単語が抽出元の文書中に２回出現していることを示す。

【００４５】次に、検索対象単語情報書込部１８が起動され、検索対象単語情報格納バッファ４２の内容（単語情報）が外部記憶装置４に格納される（ステップ３０６）。この後、ステップ３０３に戻り、外部記憶装置４に格納された次の検索対象文書を読み出し、その検索対象文書からの単語の抽出と出現頻度の算出を行う。このようにして外部記憶装置４に格納されたすべての検索対

象文書について単語の抽出および出現頻度の算出を行い、その結果を外部記憶装置４に格納する。

【００４６】外部記憶装置４に格納されたすべての検索対象文書の単語情報が外部記憶装置４に格納されたら、次に全検索対象単語統計算出部１９が起動される。全検索対象単語統計算出部１９は、外部記憶装置４に格納された全検索対象文書の単語情報（出現頻度の情報）を順次読み出して検索対象文書格納バッファ４１に格納し、この検索対象文書格納バッファ４１に格納された、全検索対象文書の単語情報（出現頻度の情報）単語の出現頻度に基づき、個々の単語の出現頻度の統計値（例えば出現文書数など）を算出する。そして、図９に示すように、このように算出された個々の単語の出現頻度の出現文書数など統計値を、全検索対象単語情報格納バッファ４３に全検索対象単語情報として格納する（ステップ３０７）。　次に、不要語リスト作成部２１が起動される。不要語リスト作成部２１は、全検索対象単語情報格納バッファ４３に格納されている全検索対象単語情報と不要語設定バッファ４４に格納された不要語設定基準に基づいて不要語リストを作成し、作成された不要語リストを外部記憶装置４に格納する（ステップ３０８）。この不要語リストの作成は、ユーザにより任意に指定された不要語の選択基準に基づいて行われる。

【００４７】不要語設定部２０にてユーザにより任意の数の単語（不要語）が指定された場合（第１の方法の場合）、不要語リスト作成部２１は、その不要語の出現頻度を全検索対象単語情報格納バッファ４３から読み出し、読み出した不要語の出現頻度の中の最小出現頻度を基準として、出現頻度が基準値以上の単語を全検索対象単語情報格納バッファ４３の中からすべて抽出し、これを不要語として不要語バッファ４５に格納する。図１０にこの不要語バッファ４５に格納された不要語の例を示す。

【００４８】また、不要語設定部２０にてユーザにより不要語の数（或いは出現頻度値の順位）が指定された場合（第２の方法の場合）、不要語リスト作成部２１は、全検索対象単語情報の中で、出現頻度が高いものから優先に指定数（指定順位）までの単語を不要語として決定して不要語バッファ４５に格納する。

【００４９】さらに、不要語設定部２０にてユーザにより出現頻度の基準値が指定された場合（第３の方法の場合）、不要語リスト作成部２１は、出現頻度が基準値以上のすべての単語を不要語として決定し、不要語バッファ４５に格納する。

【００５０】以上により、文書データベースおよび不要語リストの作成が終了する。

【００５１】続いて、類似文書検索の動作について図１１乃至図１６を参照して説明する。まず、初期化部１２が起動され、全バッファの初期化が行われる（ステップ４０１）。次に、不要語リスト読出部２２が起動され、

外部記憶装置４から不要語リストを読み出して不要語バッファ４５に格納する（ステップ４０２）。

【００５２】次に、検索キー文書入力部２３が起動されることで、ユーザにより指定された検索キー文書が外部記憶装置４から読み込まれ、読み込まれた検索キー文書が検索キー文書格納バッファ４６に格納される（ステップ４０３）。図１２に検索キー文書格納バッファ４６に格納された検索キー文書の例を示す。

【００５３】続いて、検索キー単語抽出部２４が起動される。検索キー単語抽出部２４は、検索キー文書格納バッファ４６に格納された検索キー文書から形態素解析等によって単語を切り出し、切り出された単語群から「名詞」や「サ変名詞」などの文書の内容を特徴付ける単語を抽出し、抽出された単語を検索キー単語情報格納バッファ４７に格納する（ステップ４０４）。

【００５４】次に、不要語除去部２８が起動される。不要語除去部２８は、検索キー単語情報格納バッファ４７に格納されている検索キー文書の単語群の中から、不要語バッファ４５に格納されている不要語と一致する単語を見つけ出してこれを削除する（ステップ４０５）。

【００５５】続いて、検索キー単語出現頻度算出部２５が起動される。検索キー単語出現頻度算出部２５は、検索キー単語情報格納バッファ４７に格納されている個々の単語について、その抽出元文書中での出現頻度を算出し、算出された出現頻度の情報を、図１３に示すように、検索キー単語情報格納バッファ４７において単語と対応付けて格納する（ステップ４０６）。

【００５６】次に、検索対象文書読出部１５が起動される。検索対象文書読出部１５は、外部記憶装置４にまだ処理を終えてない検索対象文書あるか否かを判断し（ステップ４０７）、もし検索対象文書があれば、その検索対象文書を検索対象文書格納バッファ４１に格納する。この後、検索対象文書単語抽出部１６によって、検索対象文書格納バッファ４１に格納された検索対象文書から形態素解析等によって単語の切り出しが行われ、切り出された単語群の中から「名詞」や「サ変名詞」などの文書の内容を特徴付ける単語が抽出され、抽出された単語の情報が検索対象単語情報格納バッファ４２に格納される（ステップ４０８）。

【００５７】続いて、不要語除去部２８が起動される。不要語除去部２８は、検索対象単語情報格納バッファ４２に格納されている検索対象文書の単語群の中から、不要語バッファ４５に格納されている不要語と一致する単語を見つけ出してこれを削除する（ステップ４０９）。

【００５８】次に、共通単語抽出部２７が起動される。共通単語抽出部２７は、それぞれ不要語の削除を終えた検索対象単語情報格納バッファ４２と検索キー単語情報格納バッファ４７内から共通に格納されている単語を検出し、図１４に示すように、その検出された単語を共通単語情報格納バッファ４８に格納する（ステップ４１０）。

【００５９】この後、類似度算出部２９が起動される。類似度算出部２９は、検索対象単語情報格納バッファ４２、検索キー単語情報格納バッファ４７および共通単語情報格納バッファ４８の内容を基に、ベクトル空間法等により、検索キー文書と検索対象文書との類似度を算出し、算出された類似度を類似度格納バッファ４９に格納する（ステップ４１１）。図１５に、この類似度格納バッファ４９に格納された検索キー文書と個々の検索対象文書との類似度情報の例を示す。

【００６０】この後、ステップ４０７に戻り、外部記憶装置４にまだ処理を終えてない検索対象文書がある場合は、その検索対象文書について前記と同様の処理を行い、こうして算出された検索キー文書と検索対象文書との類似度を類似度格納バッファ４９に格納する。

【００６１】外部記憶装置４に格納されたすべての検索対象文書と検索キー文書との類似度が類似度格納バッファ４９に格納された後、検索結果出力部３０が起動される。検索結果出力部３０は、類似度格納バッファ４９の内容から、例えば図１６に示すように、類似度が高いものから順に検索対象文書のＩＤを並べ、その結果を検索結果出力バッファ５０に格納する。この後、出力部１４によって、検索結果出力バッファ５０の内容が表示装置３に出力される（ステップ４１２）。

【００６２】かくして本実施形態の類似文書検索装置によれば、不要語リストの作成する際のユーザの作業負荷が大幅に軽減され、全般的な類似文書検索の効率アップを図ることができる。すなわち、本実施形態の類似文書検索装置において、不要語リストを作成するために必要となるユーザの操作は、１つ乃至少数の不要語の指定、或いは不要語の数の指定、或いは出現頻度の基準値のいずれかでよく、このような簡単な指定操作がユーザによって事前に行われるだけで、最適かつ妥当な不要語を漏れなくリストアップでき、類似文書検索の精度の向上と安定化を図ることができる。

【００６３】なお、本実施形態では、不要語の代表としてユーザにより指定された単語の出現頻度を基準値として、この基準値以上の出現頻度を持つすべての単語を不要語として設定する場合について説明したが、この基準値よりも低値側に一定マージンを確保して、このマージン内の出現頻度をもつ単語も不要語として判定するようにしてもよい。

【００６４】

【発明の効果】以上説明したように本発明によれば、不要語リストの作成においてユーザが不要語とすべき単語を一つ一つ登録しなくてもよく、例えば、ユーザが１つ乃至少数の不要語を指定したり、不要語の数を指定したり、出現頻度の基準値を入力するだけで、所望の不要語リストを作成することができる。これにより、類似文書検索の全体的な効率を高めることができ、また、検索対

13

象文書データベースの種類毎に最適かつ妥当な不要語を漏れなく抽出することができるので、類似文書検索の精度の向上と安定化を図ることができる。

【図面の簡単な説明】

【図１】本発明の実施形態である類似文書検索装置のハードウェア構成を示す図

【図２】図１の類似文書検索装置の制御装置の構成を示す機能ブロック図

【図３】文書データベースおよび不要語リストの作成手順を示すフローチャート

【図４】ユーザにより指定された不要語の例を示す図

【図５】ユーザにより指定された不要語の数（出現頻度の順位）の例を示す図

【図６】ユーザにより指定された出現頻度の基準値の例を示す図

【図７】検索対象文書の例を示す図

【図８】検索対象単語情報格納バッファに格納された単語と出現頻度の例を示す図

【図９】全検索対象単語情報格納バッファに格納された、全検索対象文書の単語とその出現頻度の統計値の例を示す図

【図１０】不要語バッファに格納された不要語の例を示す図

【図１１】類似文書検索の動作の手順を示すフローチャート

【図１２】検索キー文書の例を示す図

【図１３】検索キー単語情報格納バッファに格納された

14

単語と出現頻度の例を示す図

【図１４】共通単語情報格納バッファに格納された共通単語と出現頻度の例を示す図

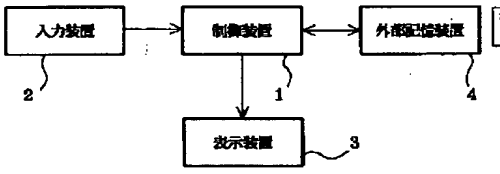
【図１５】類似度格納バッファに格納された検索キー文書と検索対象文書との類似度の例を示す図

【図１６】類似文書検索結果の例を示す図

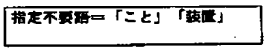
【符号の説明】

２００……メイン処理部
２０１……初期化部
２０２……入力部
２０３……出力部
２０４……検索対象文書読み出し部
２０５……検索対象文書単語抽出部
２０６……検索対象単語出現頻度算出部
２０７……検索対象単語情報算出部
２０８……検索対象単語統計算出部
２０９……不要語設定部
２１０……不要語リスト作成部
２１１……不要語リスト読み出し部
２１２……検索キー文書入力部
２１３……検索キー単語抽出部
２１４……検索キー単語出現頻度算出部
２１５……検索対象単語情報読み出し部
２１６……共通単語抽出部
２１７……不要語除去部
２１８……類似度算出部
２１９……検索結果出力部

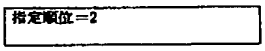
【図１】



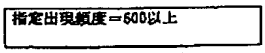
【図４】



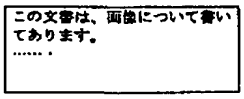
【図５】



【図６】



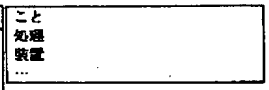
【図７】



【図８】

単語	頻度
文書	2
画像	3
……	……

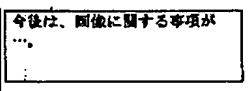
【図１０】



【図９】

順位	単語	出現頻度（例えば出現文書数など）
1	こと	627
2	処理	480
3	装置	420
4	カラー	100
…	…	…

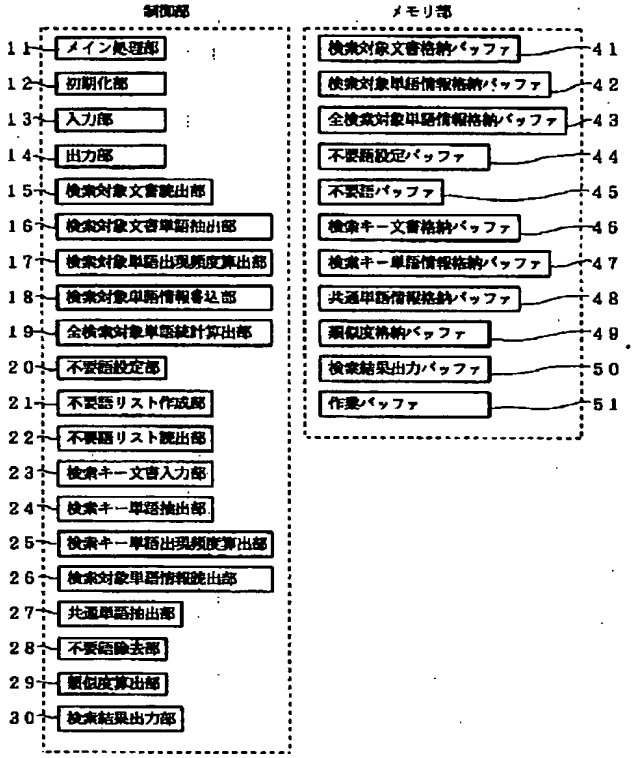
【図１２】



【図１３】

単語	頻度
今後	2
画像	3
……	……

【図２】



【図１４】

単語	頻度
画像	3
……	……

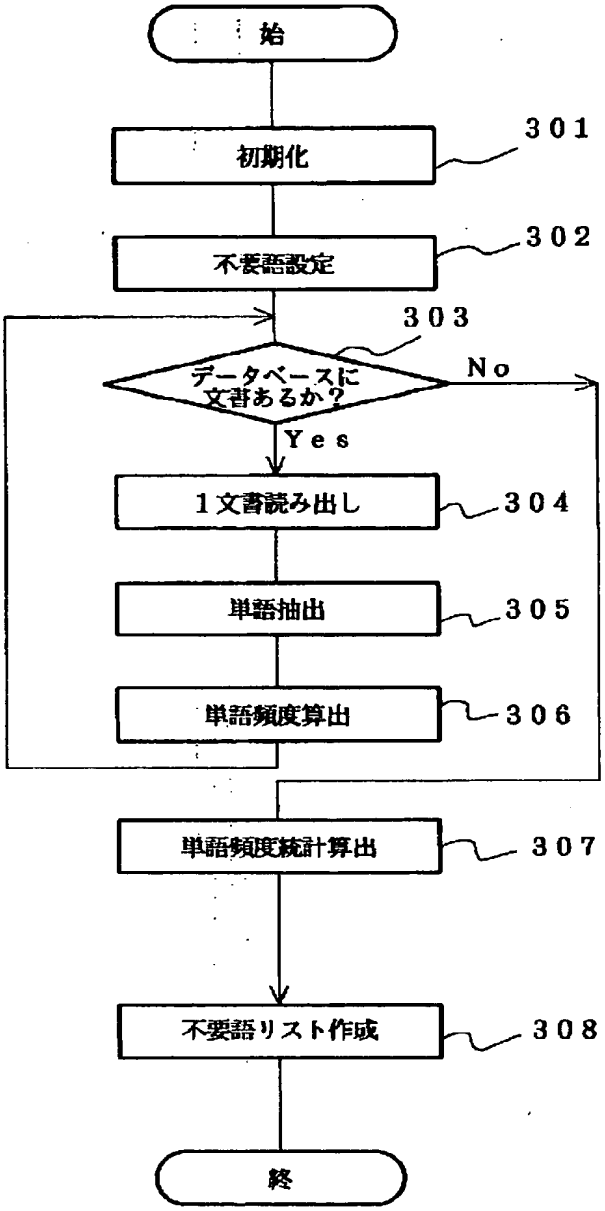
【図１５】

検索対象文書ID	類似度
1	0.2454
2	0.5842
3	0.9542
……	……

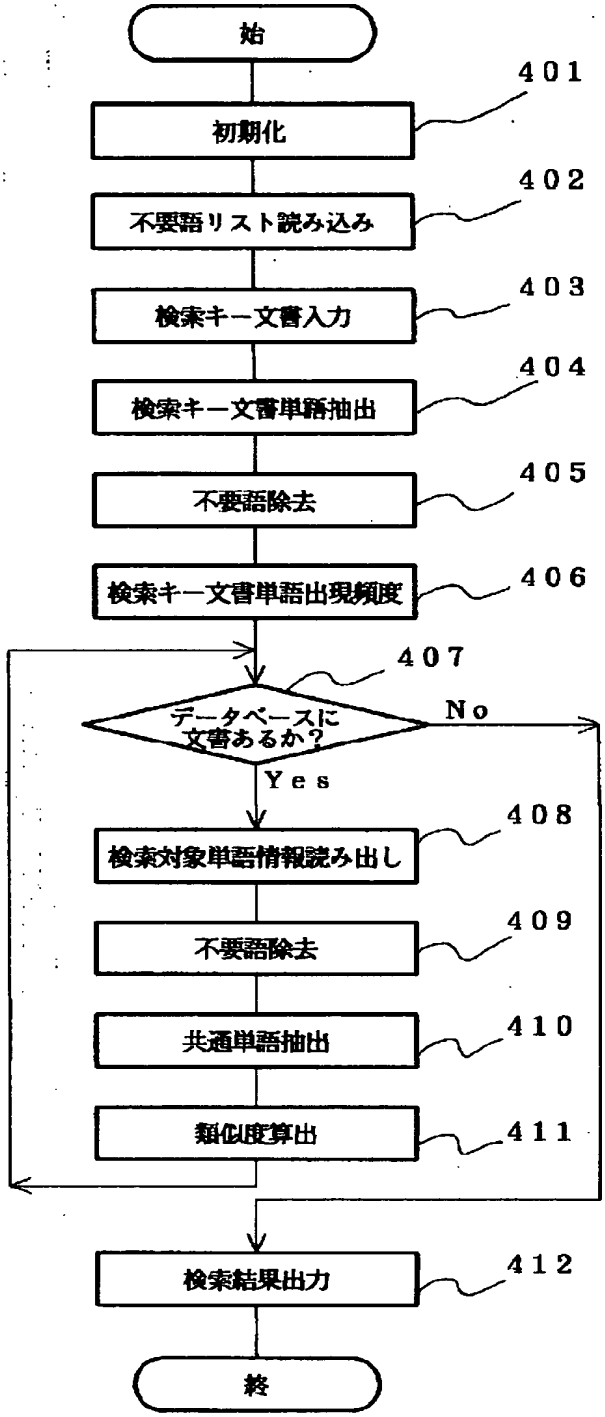
【図１６】

類似検索対象文書＝
42
54
314
……

【図３】


始
初期化
301
不要語設定
302
303
データベースに文書あるか？
No
Yes
１文書読み出し
304
単語抽出
305
単語頻度算出
306
単語頻度統計算出
307
不要語リスト作成
308
終

【図１１】


始
401
初期化
402
不要語リスト読み込み
403
検索キー文書入力
404
検索キー文書単語抽出
405
不要語除去
406
検索キー文書単語出現頻度
407
データベースに文書あるか？
No
Yes
408
検索対象単語情報読み出し
409
不要語除去
410
共通単語抽出
411
類似度算出
412
検索結果出力
終

フロントページの続き

(72)発明者　中本　幸夫
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　仁科　卓哉
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　久保田　直秀
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内